Journal of Informatics and Interactive Technolo
Volume 3, Number 1, April 2026
DOI : 10.63547/jiite.v3i1.141
e-ISSN : 3048-1023

Classifying MOOC Students Using k-Nearest Neighbors and Decision Trees for Early Detection of At-Risk Learners

Mohd Hafizan bin Musa[1], Sazilah binti Salam[2], Mohd Adili bin Norasikin[3], Asniyani Nur Haidar binti Abdullah[4], Azizul Mohd Yusoff[5], Uning Lestari[6]

[1,2,3,4]Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka
[1]Fakulti Sains Komputer dan Matematik, Universiti Teknologi MARA Cawangan Johor, Kampus Segamat
[2]Faculty of Engineering and Physical Sciences, University of Southampton, United Kingdom
[5]Kolej Komuniti Segamat, Johor, Malaysia
[6]Faculty of Science and Information Technology, Universitas Akprind Yogyakarta, Indonesia
email:[1]p032220014@student.utem.edu.my, mohdh233@uitm.edu.my, [2]sazilah@utem.edu.my, S.Binti-Salam@soton.ac.uk, [3]adili@utem.edu.my, [4]asniyani@utem.edu.my, [5]azizulkksl@gmail.com, [6]uning@akprind.ac.id

Article Info

Article history:
Received 20-02-2026
Revised 02-03-2026
Received 17-03-2026

Keywords:
At-Risk Student
MOOC
Machine Learning
Classification
Decision Tree
K-Nearest Neighbor

ABSTRACT

Massive Open Online Courses (MOOCs) have expanded access to education, but they consistently face issues with a high dropout rate. At-risk students need to be identified early so that they can be promptly intervened upon and achieve better academic results. This study investigates the use of supervised machine learning, specifically Decision Trees (DT) and k-Nearest Neighbors (kNN), to classify MOOC students from Institution A based on behavioural and engagement indicators. The predictive features include the number of comments posted in discussion forums, the number of kudos received from other learners, overall course progress, and time spent on learning pages. These variables represent both social interaction and individual engagement dimensions of learning behaviour. A 552 dataset of MOOC participants was pre-processed and analysed, followed by classification experiments using DT and kNN. Model performance was evaluated using accuracy, precision, recall, and F1-score. Evaluation results showed that the kNN algorithm delivered the best performance, achieving a score of 99.46% accuracy, 99.48% precision, 99.46% recall, and a 99.43% F1-score. Meanwhile, the kNN algorithm achieved 98.73% accuracy, 98.83% precision, 98.73% recall, and a 98.72% F1. Findings suggest that the kNN model is more effective for classifying at-risk learners and can be utilised as a decision support tool to identify such learners.

Corresponding author:
Sazilah binti Salam
Fakulti Teknologi Maklumat dan Komunikasi, Universiti Teknikal Malaysia Melaka
Email: sazilah@utem.edu.my

1. INTRODUCTION

In helping the institution in managing their teaching and learning process to become smooth, multiple e-learning platforms are optimised, and this includes Learning Management Systems (LMS), Student Information Systems (SIS) and Massive Open Online Courses (MOOC). The uses of each platform are different and present

different advantages [1], [2]. The main aim of SISs is to maintain student records and academic information so that the administration is not complicated. LMSs, in turn, enable teaching and learning actions by means of the Moodle, Google Classroom, and micro-credential systems, the main emphasis being placed on full-time students. In the meantime, MOOCs have been used to open courses to a broader audience, including both full-time and part-time learners.

A MOOC is usually made up of more structured learning materials in the form of lecture videos, interactive multimedia learning, reading materials, self-assessment assignments, discussion forums, and peer-to-peer interactions that promote active learning and collaborative knowledge building [3]. These characteristics enable learners not only to interact with the course content but also with instructors and other participants, contributing to the formation of a sense of community in a virtual classroom. MOOCs are widely recognised for promoting flexibility and self-paced learning, making them suitable for both formal academic use and lifelong learning [4]. As highlighted by North et al. (2021) [5], MOOCs are a groundbreaking educational model that can be scaled, accessible, and interactive, providing learners with an opportunity to access higher education beyond the physical limits of traditional institutions.

Although promising, MOOCs still face high dropout rates, with a very small percentage of their enrolled learners completing their courses [6]. In some cases, the dropout rate can reach 90% for certain courses that are offered by the institution [6]. Coffrin et al., (2014) [7] state that the MOOC is called Principles of Macroeconomics, and it is provided by the University of Melbourne, where 54,217 students have enrolled, with 32,598 successfully attending the course and only 1,412 having fully finished the course and gotten the certificate (4.33%). Similarly, a broader study by Jordan (2014) examined 91 MOOCs, where enrolments ranged from 450 to 226,652 students (with an average of 43,000). The findings showed that most courses had completion rates below 10%, with the overall average at just 6.5% [8]. Some of the reasons this happens are due to a combination of learner-focused factors, like lack of motivation, time constraints, and insufficient prior knowledge, and MOOC-related factors, such as isolating course structures, limited interaction, unclear design, and hidden costs [9]. This enduring problem has prompted studies on early identification of at-risk students so that teachers and administrators can develop timely interventions that have the potential to enhance retention and overall student outcomes.

Student motivation has been identified as one of the strongest predictors of MOOC completion and supported by several studies conducted by [9], [10], [11]. Engagement and behavioural measures can be used to assess student motivation in terms of participation in forums, course material development, and the duration required to complete learning tasks. As previous research has shown, active learners who engage in peer interactions, consistently invest time in the course material, and ensure progress are more likely to be successful. However, engagement patterns vary across institutions and platforms, which underscores the importance of exploring context-specific indicators that may strengthen predictive models.

At Institution A, the MOOC platform incorporates a unique "Kudos" feature, where learners or lecturers can give acknowledgement and encourage student contributions throughout the learning process. For example, if the learner provides a detailed answer to any of the activity questions and the answers appear to be very helpful for other learners to understand the topic, not only the lecturers but also other learners can give Kudos to the learner. Directly, this will encourage more students to excel in the MOOC session and collect as many kudos as possible. This aspect is another dimension of instructor-student interaction that can have a positive effect on motivation and perseverance. The indicator, when combined with other engagement measures, such as the number of comments on forums, course progress, and time spent on learning pages, will provide a more profound insight into how students behave and what dropout predictors may be.

To overcome the issue of early dropout detection, the proposed study will utilize supervised machine learning, specifically Decision Trees (DT) and k-Nearest Neighbours (kNN), to categorize MOOC students at Institution A as either completers or non-completers. The algorithms can effectively fit educational data analysis due to their interpretability and suitability in addressing classification problems. The paper also compares the predictive capacity of the various engagement indicators and the performance of the two algorithms in forecasting at-risk students. With an emphasis on the MOOC platform of Institution A, this study makes a methodological contribution through the application of machine learning to early warning systems and a practical contribution by offering insights that can guide institutional-level student retention strategies.

The following sections of this article are structured as follows: Section 2 reviews existing classification-based machine learning techniques and outlines the motivations derived from prior studies aimed at predicting student performance. Section 3 provides a detailed description of the research design and experimental setup. Section 4 presents the empirical findings, and Section 5 concludes the study with key insights and implications.

2. LITERATURE REVIEW

2.1. Classification algorithm in the higher education field

The use of classification algorithms in the educational sector has undergone massive application by both data mining and machine learning systems. Cases covered by this study include forecasting graduation rates, the identification of at-risk students, the categorisation of students by interest or aptitude as a means of recommending appropriate electives or specialisations, and even anticipating students who are most likely to switch majors or drop out of the program as a way of boosting retention efforts.

For instance, [12] developed an ensemble classification model using both synchronous and asynchronous online learning behaviour to predict students at risk of academic failure during the COVID-19 pandemic. Their model, which included the number of lecture materials downloaded, attendance, quiz score, and online session time, achieved a specificity of about 90.34%. In a different study by [13], a proprietary classification framework was developed, combining three algorithms with a random forest–based predictive model to forecast student performance and identify at-risk students in introductory programming courses. In contrast, [14] employed five established algorithms—Logistic Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), k-Nearest Neighbours (KNN), and Naïve Bayes (NB) to construct predictive models of student performance. These multi-class models classified students into three categories: "likely to fail" (grade < 45%), "borderline" (45% ≤ grade ≤ 55%), and "likely to pass" (grade > 55%). Notably, the final examination mark, weighted at 40%, was excluded from the models on the grounds that it occurred at the end of the academic term and thus would not allow for timely interventions. A similar structure of classification approaches, based on specific algorithms, is also evident in the studies conducted by [15], [16].

Another study by [17] examined student attrition among upper-year physiotherapy students using 23 supervised classifiers; they obtained best results with the Subspace k-NN algorithm, reaching an accuracy of approximately 86.3%. In addition, research on elective course selection by [18] explored how student attributes influence choices of electives, providing insights useful for classifying students by interest/aptitude for better elective recommendations.

In this study, Decision Tree (DT) and k-Nearest Neighbours (kNN) are employed because they provide greater simplicity, interpretability, and are especially well suited to small-to-medium sized datasets, enabling clear decision rules and minimal parameter tuning. These methods make it easier to understand how classifications are made and to implement timely interventions. While more complex algorithms such as Random Forest (RF) and Support Vector Machine (SVM) often yield stronger predictive performance under many settings, they tend to be less transparent and require more computational resources and careful parameterisation. For example, a recent comparative study of classification algorithms found that Decision Trees and kNN are among the most frequently used interpretable models, due to their inherent intelligibility, whereas RF, SVM and other ensemble or margin-based methods dominate in performance when interpretability is less of a constraint[19].

3. METHODOLOGY

Figure 1 illustrates the research framework, which begins with the collection of MOOC engagement profiles from a dataset of 552 students enrolled in a single semester of a Mandarin course. The process involves data preprocessing and feature selection, followed by splitting the data into training and testing subsets. Both kNN and DT algorithms are applied for classification, after which the models are evaluated for their predictive performance. This research design aligns with established methodologies utilized in previous studies [20], [21], [22]. Finally, the findings are interpreted to provide recommendations for improving learning outcomes for at-risk learners.

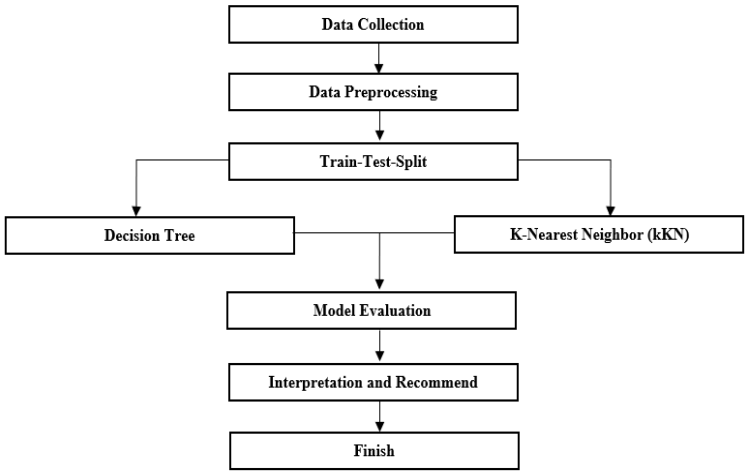

Figure 1.Research design flow

The study begins with data collection to identify at-risk students based on behavioral and engagement indicators. These predictive features include the number of forum comments, kudos received, course progress, and time spent on learning pages. During Data Preprocessing, missing values and inconsistencies are handled, and categorical data are encoded. For example, the 'time spent' feature is aggregated into hours, and the 'certificate_id' is converted into a binary integer (0 = non-completer, 1 = completer).

Subsequently, the dataset is divided into training and testing subsets during the Train-Test-Split stage, typically following a 70:30 ratio. This proportion is commonly employed in machine learning research to ensure a balanced trade-off between effective model training and reliable performance evaluation. Allocating approximately 70% of the data for training enables the model to capture essential patterns within the dataset, while the remaining 30% is used to validate its generalisation capability on unseen data. Although the 70:30 ratio is not a strict rule, empirical studies have shown that the training data must comprise more than 70% to provide stable and consistent results across various classification problems [23]. Several studies on MOOCs have likewise employed the 70:30 data split ratio for model training and testing [24], [25].

A fixed random_state was used to ensure consistent replication of the result. For the model building stage, there are two classification algorithms implemented in parallel, namely DT and kNN. The DT builds an algorithm based on the features, splitting the data into rules recursively according to the values of the features, whereas kNN assigns the individual data point to a category according to the majority category in its k closest points in the feature space.

In addition to the initial static data partitioning, a five-fold cross-validation procedure was implemented to rigorously evaluate the model's robustness. Under this procedure, the training dataset was systematically divided into five mutually exclusive and approximately equal subsets (folds). The model was iteratively trained and validated across each fold, and the resulting performance metrics were aggregated to yield a more stable and generalizable estimate. This cross-validation strategy serves to identify and mitigate potential overfitting, thereby ensuring that the model's performance is not contingent upon any single data partition.

Finally, the model was evaluated using four key performance metrics: accuracy, precision, recall, and F1-score. These metrics are computed based on True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) values [26] and collectively provide a robust evaluation of each algorithm's capability to classify potential at-risk learners.

4. RESULTS AND DISCUSSION

4.1. Data set distribution

Figure 2 presents the box plot that shows the distribution of the relationship of all predictive variables towards the MOOC completion status. The boxplots reveal clear differences in the distributions of Comments, Progress, Kudos, and Hours spent (HoursOnly) between learners who completed the course and those who did not.

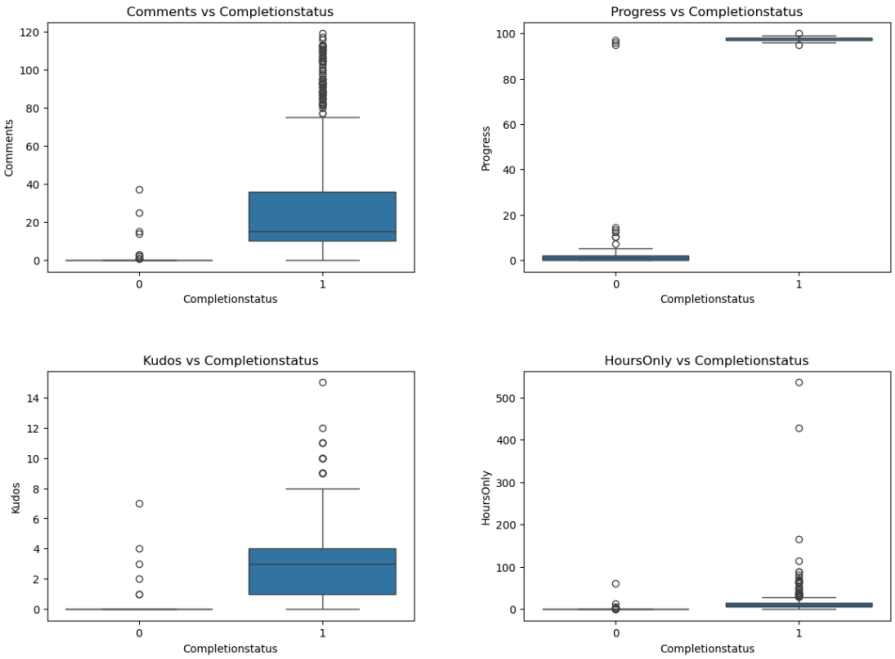

Figure 2.Comparison of comments, progress, kudos, and time spent between completion status

For all four variables, the median values for the completion group (Completionstatus = 1) are noticeably higher than those of the non-completion group (Completionstatus = 0). This trend implies that more engaged learners, who leave more comments, make more progress, have more kudos, and spend more time on the course, are more likely to complete the MOOC. The scattering of the data for the completion group is also broader, particularly for Comments and HoursOnly, indicating a wide range of learner behaviours even among those who successfully finished the course.

The presence of numerous outliers, especially in the completion group, highlights that while higher engagement generally correlates with completion, some learners exhibit extreme levels of participation. For example, several learners spent substantially more time or gave significantly more comments and kudos compared to the majority. These outliers may represent highly motivated learners or those facing unusual circumstances, such as revisiting content extensively. On the other hand, the non-completion group is much less varied, with the majority of learners grouped around minimum values of engagement measures. Overall, the distribution suggests that while consistent engagement strongly predicts course completion, the diversity of behaviours among completers underscores the importance of considering individual learning styles and contexts when interpreting MOOC participation data.

4.2. Classification model result

This research was conducted using Jupyter Notebook on the Anaconda platform. Table 1 below presents the results of both kNN and DT with k = 5.

Table 1. Comparison of performance metrics between kNN and DT

Algorithm	Accuracy	Precision	Recall	F1-Score
kNN	0.9946	0.9948	0.9946	0.9943
DT	0.9873	0.9883	0.9873	0.9872

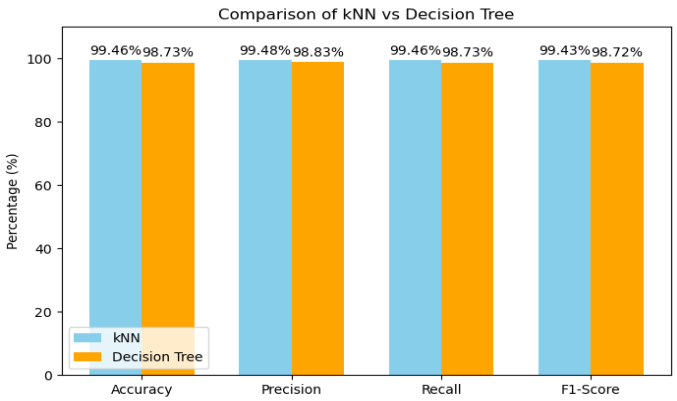

Figure 3.Illustration of comparison of performance metrics between algorithm tested

The k-Nearest Neighbors (kNN) model demonstrated slightly superior performance compared to the Decision Tree (DT) model across all evaluated metrics, achieving an overall accuracy of 99.46%. This high level of accuracy indicates the model's strong capability to correctly classify the data. In terms of precision, the kNN model achieved 99.48%, signifying that more than 99% of its positive predictions were correct. The recall score, also 99.46%, reflects the model's excellent ability to identify actual positive instances, thereby demonstrating a high degree of sensitivity. Consequently, the F1-score for kNN reached 99.43%, confirming a balanced trade-off between precision and recall. By contrast, the DT model achieved an accuracy of 98.73%, with a precision of 98.83%, recall of 98.73%, and an F1-score of 98.72%, which, although strong, remains slightly below the performance of kNN.

The exceptional accuracy (99.46%) can be attributed to the high predictive power of the 'Progress' and 'Kudos' features. As shown in the box plots (Figure 2), there is a nearly absolute separation in 'Progress' values between completers and non-completers, with successful learners consistently maintaining high engagement levels. Because these behavioral markers are such strong indicators of the target variable, the models were able to achieve high classification performance without evidence of overfitting.

The performance of the kNN model (99.46%) and the DT model (98.73%) is highly competitive when compared with results in recent literature. For instance, [17] reported an accuracy of 86.3% using a Subspace kNN algorithm to predict student attrition, while [12] achieved a specificity of 90.34% in identifying at-risk students during the COVID-19 pandemic. The high accuracy observed in this study may be attributed to the inclusion of the unique 'Kudos' feature and specific engagement indicators such as course progress and time spent on learning pages. These features appear to be highly sensitive predictors for the Mandarin course cohort, supporting the idea that context-specific indicators can significantly improve predictive models.

5. CONCLUSION

This study compared the two popular classification algorithms, k-Nearest Neighbors (kNN) and Decision Tree (DT) to classify at-risk learners during a MOOC session using a dataset of 552 records on the Mandarin course at Institution A. The results demonstrated that kNN consistently outperformed DT across all key performance metrics, achieving accuracy, precision, recall, and F1-scores of approximately 99.4%, compared to DT's 98.7% on the same measures. These results demonstrate that kNN offers a more quality and sensitive method of identifying the at-risk learners so that more and earlier interventions can be implemented to enhance learning. In addition to the numerical findings, this paper highlights the need to choose the right algorithm when precision and recall are paramount, particularly in the field of education analytics where a timely and accurate classification can result in the provision of better resources, increased support to students, and lower dropout.

K-Fold Cross-Validation was used to improve the strength of the models and reduce overfitting during the training process. This method provided more reliable estimates of performance than use of one fixed data split. Nevertheless, despite these promising results, the model was ultimately evaluated on the same dataset used for both training and validation.

Future research may extend this analysis by testing additional algorithms, applying feature selection techniques, or evaluating the models on larger and more diverse datasets to further validate the generalizability of the

findings. Moreover, the present study was constrained by the relatively small sample size; subsequent studies should consider utilizing substantially larger datasets drawn from multiple courses and academic terms to improve model robustness and support more comprehensive analyses.

ACKNOWLEDGEMENT

This research was conducted under the Pervasive Computing & Educational Technology (PET) Research Group, Centre for Advanced Computing Technology (CACT), Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM) and in collaboration with Fakulti Sains Komputer dan Matematik (FSKM), Universiti Teknologi MARA, and Web Science Institute, University of Southampton, United Kingdom.

REFERENCE

[1] S. Salah and M. Thabet, "E-Learning Management Systems: A Feature-Based Comparative Analysis," *Journal of Information Systems and Technology Management*, vol. 18, no. 0, 2021, doi: 10.4301/S1807-1775202118003.

[2] M. H. Musa, S. Salam, M. A. Norasikin, S. Shabarudin, I. Ahmad, and W. S. N. Saifudin, "A Comprehensive Review of Data Retrieval and Evaluation Methods for Current Universities Ontological Model," in *Proc. 16th Int. Conf. Knowledge and System Engineering (KSE)*, 2024, pp. 205–212, doi: 10.1109/KSE63888.2024.11063480.

[3] F. C. Bonafini, "The Effects of Participants' Engagement with Videos and Forums in a MOOC for Teachers' Professional Development," *Open Praxis*, vol. 9, no. 4, pp. 433–447, 2017.

[4] A. M. Anson, "Beyond the Classroom: MOOCs and the Evolution of Lifelong Learning," *Journal of Computer Science Application and Engineering (JOSAPEN)*, vol. 2, no. 1, pp. 6–10, 2024.

[5] S. M. North, R. Richardson, and M. M. North, "To Adapt MOOCs, or Not? That Is No Longer the Question," *Universal Journal of Educational Research*, vol. 2, no. 1, pp. 69–72, 2014.

[6] J. Zhang, M. Gao, and J. Zhang, "The Learning Behaviours of Dropouts in MOOCs: A Collective Attention Network Perspective," *Computers & Education*, vol. 167, p. 104189, 2021, doi: 10.1016/j.compedu.2021.104189.

[7] C. Coffrin, L. Corrin, P. de Barba, and G. Kennedy, "Visualizing Patterns of Student Engagement and Performance in MOOCs," in *Proc. ACM Conf.*, 2014, doi: 10.1145/2567574.2567586.

[8] K. Jordan, "Initial Trends in Enrolment and Completion of Massive Open Online Courses," *International Review of Research in Open and Distance Learning*, vol. 15, no. 1, pp. 133–160, 2014, doi: 10.19173/irrodl.v15i1.1651.

[9] H. Khalil and M. Ebner, "MOOCs Completion Rates and Possible Methods to Improve Retention: A Literature Review," in *Proc. World Conf. Educational Multimedia, Hypermedia and Telecommunications*, 2014, pp. 1305–1313.

[10] K. A. Azhar, N. Iqbal, Z. Shah, and H. Ahmed, "Understanding High Dropout Rates in MOOCs: A Qualitative Case Study from Pakistan," *Innovations in Education and Teaching International*, vol. 61, no. 4, pp. 764–778, 2024, doi: 10.1080/14703297.2023.2200753.

[11] L. N. Bezerra and M. T. Silva, "A Review of Literature on the Reasons That Cause the High Dropout Rates in MOOCs," *Revista Espacios*, vol. 38, no. 5, 2017.

[12] H. Karalar, C. Kapucu, and H. Gürüler, "Predicting Students at Risk of Academic Failure Using Ensemble Model During Pandemic in a Distance Learning System," *International Journal of Educational Technology in Higher Education*, vol. 18, no. 1, p. 63, 2021, doi: 10.1186/s41239-021-00300-y.

[13] A. Kumar Veerasamy, D. D'Souza, M.-V. Apiola, M.-J. Laakso, and T. Salakoski, "Using Early Assessment Performance as Early Warning Signs to Identify At-Risk Students in Programming Courses," in *Proc. IEEE Frontiers in Education Conf. (FIE)*, 2020, pp. 1–9, doi: 10.1109/FIE44824.2020.9274277.

[14] K. Alalawi, R. Athauda, R. Chiong, and I. Renner, "Evaluating the Student Performance Prediction and Action Framework Through a Learning Analytics Intervention Study," *Education and Information Technologies*, 2024, doi: 10.1007/s10639-024-12923-5.

[15] N. Roslan, J. Mohd Jamil, I. N. Mohd Shaharanee, and S. Alalawi, "Prediction of Student Dropout in Malaysian Private Higher Education Institute Using Data Mining Application," *Journal of Advanced Research in Applied Sciences and Engineering Technology*, vol. 45, pp. 168–176, 2024, doi: 10.37934/araset.45.2.168176.

[16] E. Nimy, M. Mosia, and C. Chibaya, "Identifying At-Risk Students for Early Intervention: A Probabilistic Machine Learning Approach," *Applied Sciences*, vol. 13, no. 6, 2023, doi: 10.3390/app13063869.

[17] M. Barramuño, C. Meza-Narváez, and G. Gálvez-García, "Prediction of Student Attrition Risk Using Machine Learning," *Journal of Applied Research in Higher Education*, vol. 14, no. 3, pp. 974–986, 2022.

[18] D. H. Ting and C. K. C. Lee, "Understanding Students' Choice of Electives and Its Implications," *Studies in Higher Education*, vol. 37, no. 3, pp. 309–325, 2012, doi: 10.1080/03075079.2010.512383.

[19] A. Jovic, N. Frid, K. Brkic, and M. Cifrek, "Interpretability and Accuracy of Machine Learning Algorithms for Biomedical Time Series Analysis: A Scoping Review," *Biomedical Signal Processing and Control*, vol. 110, p. 108153, 2025, doi: 10.1016/j.bspc.2025.108153.

[20] S. Wiyono, D. S. Wibowo, M. F. Hidayatullah, and D. Dairoh, "Comparative Study of KNN, SVM and Decision Tree Algorithm for Student Performance Prediction," *International Journal of Computing Science and Applied Mathematics*, vol. 6, no. 2, pp. 50–53, 2020.

[21] H. A. Althibyani, "Predicting Student Success in MOOCs: A Comprehensive Analysis Using Machine Learning Models," *PeerJ Computer Science*, vol. 10, p. e2221, 2024.

[22] Z. Chi, S. Zhang, and L. Shi, "Analysis and Prediction of MOOC Learners' Dropout Behavior," *Applied Sciences*, vol. 13, no. 2, 2023, doi:

10.3390/app13021068.
[23] H. Bichri, A. Chergui, and M. Hain, “Investigating the Impact of Train/Test Split Ratio on the Performance of Pre-Trained Models with Custom Datasets,” *International Journal of Advanced Computer Science & Applications*, vol. 15, no. 2, 2024.
[24] N. Mansouri, M. Soui, and M. Abed, “SFS Feature Selection with Decision Tree Classifier for Massive Open Online Courses Recommendation,” *Journal of Computers in Education*, vol. 11, no. 4, pp. 1089–1110, 2024.
[25] F. A. Al-Azazi and M. Ghurab, “ANN-LSTM: A Deep Learning Model for Early Student Performance Prediction in MOOC,” *Heliyon*, vol. 9, no. 4, 2023.
[26] A. A. Septa, A. Al Farizi, A. N. Khafid, D. Prasetyo, N. C. Romadhon, and F. S. Utomo, “Diabetes Detection Optimisation with Hyperparameter Tuning in Random Forest Algorithm,” *Journal of Informatics and Interactive Technology*, vol. 1, no. 3, pp. 165–177, 2024.